DOCUMENT RESUME

ED 432 588 TM 029 964

AUTHOR Sundre, Donna L.
TITLE Does Examinee Motivation Moderate the Relationship between Test Consequences and Test Performance?
PUB DATE 1999-04-00
NOTE 21p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).
PUB TYPE Numerical/Quantitative Data (110) -- Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Effect Size; Higher Education; *Performance Factors; Student Attitudes; *Student Motivation; Tables (Data); *Undergraduate Students
IDENTIFIERS *Consequences; *High Stakes Tests

ABSTRACT

There is substantial evidence that the disposition of test takers is central to performance. This research extends previous work by replicating the experimental design of L. Wolf and J. Smith (1995) and conducting a secondary analysis of their data to attempt to demonstrate differential effect sizes for examinees reporting varying motivation levels in consequences versus no-consequences situations. The second phase of the study investigated the hypothesis that examinee motivation and item type moderate the relationship between test consequences and test performance by expanding the Wolf and Smith design to include multiple-choice and essay test items. Ninety undergraduate psychology majors participated in the study. The study supported previous findings concerning increases in self-reports of motivation and enhanced test performances in testing situations with consequences. The motivation subscores of Importance and Effort were found to be useful, and for some of the hypotheses the two scores behaved differently and in accord with expectations concerning the construct and context. The effect sizes observed for motivation and performance in the essay format compared to the multiple-choice were among the largest obtained in the study. (Contains 2 figures, 14 tables, and 9 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 432 588

Does Examinee Motivation Moderate the Relationship between Test Consequences and Test Performance?

Donna L. Sundre
Center for Assessment and Research Studies
James Madison University

A symposium paper session presented at the AERA meeting
Montreal, Canada
April, 1999

TM029964

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
Donna Sundre
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

BEST COPY AVAILABLE

Introduction

Many researchers have explored the relationship between examinee motivation and test performances. A number of studies have been conducted comparing student test taking motivation with paper and pencil multiple-choice examinations in consequence and no-consequence test conditions (Burke, 1991; Wolf and Smith, 1995; Wolf, Smith, & Birnbaum, 1994). Kim and McLean demonstrated significant score increases using special instructions in a CAT testing administration. A few studies have attempted to establish motivation levels for different content areas (Sundre, 1997). Wolf, Smith, & Birnbaum (1994) explored the impact of the difficulty of a task and demonstrated that 'arduousness' requires greater motivation to expend the effort to complete or correctly perform a difficult task. Their work demonstrated that lack of motivation negatively influences test performance above and beyond known ability levels. Zeidner (1993) showed that students prefer multiple-choice items to more challenging and demanding performance assessments. However, Herman, Davina, and Wakai (1997) reported that students indicated they try harder on alternative-assessment items, though they don't necessarily like the challenges. Students reported that multiple-choice items are easier to understand and that they believe they perform better on them. It has been established that examinee motivation varies systematically across different contexts and tasks. It has also been shown that motivation does seem to impact performance. Accepting the negative effects of low motivation, an additional strand of studies has attempted to influence examinee motivation using various manipulations such as special test instructions (Brown & Walberg, 1993; Kim & McLean, 1995), or financial incentives (Kiplinger & Linn, 1993). Others have simply referred to the motivation concern as 'intractable' (Ewell, 1991), suggesting that assessments will become increasingly naturalistic by embedding them within course activities.

There is substantial evidence indicating that the disposition of test-takers is central to performance (Schmidt & Ryan, 1992). The research presented in this paper extends the work of previous efforts by replicating the experimental design of Wolf and Smith (1995) and conducting a secondary analysis of their data in an attempt to demonstrate differential effect sizes for examinees reporting varying motivation levels in

consequences vs. no-consequences conditions. Building upon the experimental findings of Wolf, Smith and Birnbaum (1995), in which the difficulty of tasks was found to impact both student motivation and test performance, this study reports on differential motivation and test performances under consequence vs. no-consequence test conditions across multiple choice and essay test items. Since constructed-response items are perceived as requiring more effort and motivation to sustain performance to completion, it is expected that both motivation and performance will be diminished for the essay assignment.

This paper presents two phases of a research study. The first phase reports a replication of Wolf and Smith's (1995) study in which college students responded to two parallel examinations under two experimental conditions. For one of these examinations, the score counted as part of the course grade (consequences), and in the other condition it did not (no consequences). Following each test administration, subjects completed an eight-item motivation scale. The test conditions were counterbalanced. Wolf and Smith reported that the examination condition with course credit consequences resulted in significantly higher reported motivation (effect size = 1.45) and test score performance (effect size = .26). The current study replicated the Wolf and Smith study with a new college sample. However, the relationship between consequences, motivation, and test performance may be a bit more complex.

More specifically, it was expected that greater decline in performance would be observed with examinees that report the greatest decrement in motivation when consequences are removed. Four subject groups were identified on the basis of self-reported motivation under the two experimental conditions. Using median splits of the motivation distributions, subjects were assigned to one of four possible groups: 1) Low-Low; 2) High-Low; 3) Low-High; and 4) High-High corresponding to their respective scores on the consequence and non-Consequence total motivation score. Effect sizes for performance on the examinations were calculated and compared for these groups. It was hypothesized that the high-high motivation group would evidence little if any effect size in test performance, while a substantial test performance effect size would be observed for examinees reporting significant motivation decline in the no-consequence condition. A parallel secondary analysis of the Wolf and Smith (1995) data was proposed, and the

authors agreed to release their data for this purpose. The results of these analyses are also reported.

The second study phase investigated the hypothesis that examinee motivation and item type moderate the relationship between test consequences and test performance. The Wolf and Smith (1995) design was expanded to include two item types: multiple-choice and essay. Following the logic of previous investigators, significantly lower reports of motivation and test performance are expected for a more arduous task, in this case an essay test format. Further, the effect sizes calculated for the essay test format are expected to be much larger than those previously reported for the multiple choice item type.

These studies contribute to understanding concerning examinee motivation, its stability across consequence conditions, samples, and item types. The exploration of effect sizes produced by examinees in consequential and non-consequential test conditions and different item types and varying motivation levels can inform practitioners and policy makers regarding the appropriateness of inferences they may wish to make.

Method

This study investigated the effects of motivation under two testing conditions: 1) consequences; and 2) no-consequences and two item types: 1) multiple choice and 2) essay. For one of the consequence conditions, the test results counted toward the course grade, and in the other it did not. Each subject first responded to 30 multiple choice items and 1 essay question in either a consequence or no consequence condition. Immediately following this testing, a parallel examination of 30 multiple choice items and 1 essay question was administered in the other consequence condition. The consequence conditions were randomly assigned and counter-balanced across the subjects. Each subject was clearly informed of the consequence assigned to each test condition. Thus, each subject completed two item types across two consequence conditions. In addition to the tests, subjects completed a 10-item Likert scale motivation questionnaire immediately following each of the four test conditions. This questionnaire measures the level of effort and importance students ascribe to each of the four assessment activities. All multiple-choice responses were placed onto machine-readable forms and scored via computer. The essays were completed in a separate blue book and

evaluated by the course instructor, who was blinded as to which essay topic was consequential in the experiment.

Subjects

Subjects were college students enrolled in an undergraduate psychology of personality course at a comprehensive state institution in Virginia. All students were university psychology majors. A total of ninety subjects were invited to participate in the study. Subjects received 5 points for their participation in the study.

Materials

The tests instruments were developed by the author and a faculty colleague, and covered the first five chapters presented in the personality psychology course. Two parallel forms of the test were created, each including 30 multiple-choice items (worth 60 points) and 1 essay item (worth 40 points). Many multiple-choice items were sampled from previous course examinations and an item bank accompanying the class textbook. The two parallel forms of the test were included in a single test booklet with the two sections of the test clearly marked "This test counts for your grade" and "This test does not count for your grade." Further, students were asked to indicate on their answer sheet an 'A' = test counts or 'B' = test does not count on the first item of each form of the test. Following each section of the test, subjects completed the 10-item motivation questionnaire. Thus, each student completed four motivation questionnaires: two for each multiple-choice examination and two for the essay portions across the two consequence conditions. The motivation questionnaire is a revised version of the Wolf and Smith (1993) instrument (see Figure 1). A primary reason for modifying the original motivation questionnaire from 8 to 10 items was to improve measurement of the two factors that have consistently emerged: Importance and Effort (Sundre, 1997). The Importance factor is comprised of items tapping the examinee's perception of the importance of the test to them (i.e., "This test was important to me."). The Effort factor assesses the level of effort and persistence the examinee provided (i.e., "I engaged in good effort throughout this test."). Several items were revised for clarity, and two items, written to assess examinee effort and task persistence, were added (see Figure 2). For the purposes of this study, the items were further modified to relate to the particular test (i.e., multiple-choice test or essay) the examinee had just completed.

Procedure

The University Institutional Review Board approved this study. The study was explained to the students during the first regularly scheduled class. The author appeared at the first class section to introduce the study to the students. The students completed consent forms on the first day of the class. On the day of the exam, the author administered the tests. The students were randomly assigned the two parallel forms of the examination. After the data was collected and the results of the examinations reported to the students, the researcher provided an extended debriefing session for the class providing additional information concerning the purposes and findings of the study.

Study Hypotheses

1. Motivation to perform is influenced by the consequence associated with test performance.
2. Motivation to perform enhances test performance.
3. A consequential testing condition will lead to a better test performance than a non-consequential testing condition.
4. Differential test performances will be observed for subjects reporting different motivation levels across test consequence conditions.
5. Modality of assessment impacts motivation and performance. Effect sizes will be largest when comparing essay performance across consequence conditions.

Results

Results for Hypothesis 1: Motivation to perform is influenced by the consequence associated with test performance.

The first set of hypothesis tests served to replicate the Wolf and Smith (1993) study using the modified motivation scale. Table 1 provides the descriptive statistics for the total motivation score under both the consequential and non-consequential test conditions as well as the dependent t-test and effect size calculations. The mean for the motivation scale when the test counted as part of the course grade was 43.11. This value is much higher than the self-reported motivation for the no-consequence test condition of 37.05. The hypothesis concerning the impact of test consequences on student self-reports of motivation was assessed with a dependent t-test. This mean difference was

Table 1. Motivation descriptive statistics, t-test, and effect size comparing test conditions

Test condition	Mean	SD	t-test	df	p	ES
Multiple Choice						
Consequence	43.11	5.52	6.37	61	.000	.79
No-Consequence	37.05	7.84				

statistically significant. In addition, the effect size was .79 standard deviations, which is smaller than the 1.45 effect size reported in the original Wolf and Smith (1993) study. Table 2 provides the descriptive statistics, t-tests, and effect sizes for the two subscores of the motivation scale. These results indicate lower motivation self-reports for both the Importance and the Effort subscores in the no-consequence test condition.

Table 2. Motivation subscore descriptive statistics, t-tests, and effect sizes comparing test conditions.

Test condition	Mean	SD	t-test	df	p	ES
Importance						
Consequence	22.13	2.73	7.66	61	.000	1.08
No-Consequence	18.37	4.21				
Effort						
Consequence	20.98	3.44	4.11	61	.000	.59
No-Consequence	18.68	4.34				

These results indicate significantly lower motivation self-reports for both the Importance and Effort subscores in the no-consequence test condition. Students provided consistently higher reports of both importance and the level of effort for the consequential test condition. The results also demonstrate some differentiation in self-reports for the two motivation subscores across test conditions. Of interest here is the substantively larger effect size obtained for self-reports on the Importance factor. In some respects this may be regarded as a validation of the self-reports since students knew that the non-consequential test condition did not count, it makes sense that they would rank the test's importance as lower. They are pragmatic, and they did pay attention.

While differences in motivation self-reports are interesting, the real issue is the extent to which actual test performance may be influenced by test consequences. This is the focus the next two sets of hypotheses.

Results of hypothesis 2: Motivation to perform enhances test performance.

The second set of hypotheses concerned the extent to which self-reports of motivation are related to test performance. These hypotheses were tested using correlation coefficients. If motivation to perform well enhances test performance, significant positive correlation coefficients should be observed between measures of motivation and performance. Table 3 provides the correlation coefficients and effect sizes describing these relationships. The correlation between motivation and the consequences test performance was not significant; only 2% of the variance was shared. However, the no-consequence test condition resulted in a significant correlation

Table 3. Correlation coefficients between motivation and test performance across testing conditions.

Test Condition	r	n	ES
Consequence	.15	62	.02
No-Consequence	.38	62	.14

coefficient that accounted for 14% of the variance in no-consequence test score performances. The correlation coefficients reported in Table 3 are markedly different than those reported by Wolf and Smith (1993). In their study, both the consequence ($r = .35$) and no-consequence ($r = .23$) motivation scores correlated significantly with test performances. In the current study, only the no-consequence condition resulted in a significant relationship. This may be due to the greater variability observed in the no-consequence condition for both the examination and for motivation scores in the current study. Wolf and Smith reported greater test performance variability in the consequence than the no-consequence condition and substantially larger variability in motivation scores in the no-consequence test condition. This enhanced variability could have contributed to the higher observed coefficients in their study. Table 4 provides parallel results for the motivation subscores, Importance and Effort across the two test conditions.

The consequence condition again provided very weak positive correlation coefficients, while the no-consequence condition resulted in significant relationships for both motivation factors that accounted for 10% and 13% of the variance respectively in test

Table 4. Correlation coefficients between motivation subscores and test performance across testing conditions.

Test Condition	r	n	ES
Importance			
Consequence	.12	62	.01
No-Consequence	.32	62	.10
Effort			
Consequence	.15	62	.02
No-Consequence	.37	62	.13

score performance. Again, these findings may in part be the result of reduced variance in test and motivation scores in the consequential condition. At a minimum, it should be noted that the test condition bearing no consequence for examinees resulted in greater variance in motivation self-reports. Further, this variance was found to be positively and significantly correlated with actual test performance. How can variability in examinee motivation affect test scores? Examinees could choose to try harder, put for the same effort, or choose not to engage fully on the task. The logical consequences of these possibilities lead to highly provocative considerations. First of all, providing best effort in a testing condition regardless of the presence of consequences cannot reasonably produce a score that will surpass the student's preparation or ability. It is quite possible for students to vary significantly on motivation to the extent that test performance could be dramatically altered. This is most likely to occur in the absence of consequences for examinees. These are important points to consider when making inferences concerning many examinations that bear little or no consequence for examinees, particularly if significant test performance differences are observed.

Trying hard is an important validity concern. The results of this set of hypotheses reveal that test performances are positively and significantly related to motivation self-

reports, but only in no-consequence test conditions. Does this significantly impact test scores? We turn to this question now.

Results of hypothesis 3. A consequential testing condition will lead to a better test performance than a non-consequential testing condition.

This set of hypotheses addressed the impact of test consequence condition on actual test performance. Table 5 presents the dependent t-test results for student test performances across the two test conditions. A significant result was obtained, and the effect size is quite large. These results are consistent with those reported by Wolf and Smith (1993). Test consequences do impact test performance. The current study

Table 5. Multiple choice test performance descriptive statistics and effect size comparing test conditions.

Test Condition	Mean	SD	t	df	p	ES
Consequence	49.93	4.88	3.54	61	.001	.62
Non-Consequence	46.13	7.47				

obtained an effect size of .62 across consequence conditions, while the Wolf and Smith study effect size was .25, considerably smaller. These results again serve to demonstrate that the disposition of examinees and the conditions of testing play an important role in the actual scores obtained. It appears that students choose to perform closer to capacity in consequence than in no-consequence conditions, even in a test condition embedded within an elective college course.

Results of hypothesis 4: Differential test performances will be observed for subjects reporting different motivation levels across test consequence conditions.

For this set of hypotheses, motivation-contrasting groups were derived from the motivation self-reports in the two experimental conditions as described in the Method section. Table 6 provides the frequencies for this new motivation variable. It is interesting that the largest group is comprised of individuals below the median in both test condition motivation self-reports. The second largest group is comprised of individuals reporting above median scores on motivation across both test conditions.

Table 7 presents group means for the consequence and no-consequence tests. When analyses of variance were conducted to determine if significant test performances would

Table 6. Motivation contrasting groups frequency distribution

Category	n	%
Low-Low	23	37.1
High-Low	11	17.7
Low-High	9	14.5
High-High	19	30.6

be observed, the results were rather striking. No significant group mean differences were observed in the consequence test condition (F (3,58) = .6436, p=.5901); however the no-

Table 7. Motivation contrasting group test performance descriptive statistics across test conditions

		Consequences		No Consequences	
Group	n	Mean	SD	Mean	SD
Low-Low	23	49.9	4.92	42.5	10.20
High-Low	11	50.9	5.82	48.0	4.29
Low-High	9	48.0	5.19	48.0	4.79
High-High	19	50.3	4.23	48.5	3.88

consequence test condition resulted in significant differences (F (3, 58) = 3.15, p = .0313). The test for homogeneity of variances was rejected for the latter ANOVA, however, this finding was ignored since the largest group had the greatest variance, rendering the ANOVA and the associated alpha more conservative than reported. Multiple comparisons suggest that the High-High motivation group significantly outperformed the Low-Low motivation group in the no-consequence test condition. These results support the contention that test consequences significantly and differentially impact test performances, but only in the no-consequence test condition. One would think that motivation might be associated with greater subject-matter interest and perhaps

better test preparation, but if this factor is strong and pervasive, why weren't differences observed in the consequence test condition? Of additional interest here for future research is the observation that the motivation scale has provided a means by which subjects may be identified and studied further.

The Wolf and Smith (1993) data was subjected to a secondary analysis to determine if the results obtained in the current study would replicate. The same procedure was used to identify motivation groups. Table 8 presents the frequencies and percentages for the motivation-contrasting groups. The two largest groups were the same

Table 8. Wolf and Smith motivation contrasting groups frequency distribution

Category	n	%
Low-Low	51	32.3
High-Low	28	17.7
Low-High	32	20.3
High-High	47	29.7

as those observed for the current study. Table 9 presents the group means for the consequence and no-consequence test conditions. The analyses of variance for both test conditions emerged as significant. The consequence test condition (F (3, 154) = 5.3542, p = .0016) follow-ups indicated that the High-High and the High-Low groups significantly outperformed the Low-Low motivation group. In the no-consequence test condition, (F (3, 154) = 8.8167, p = .0000), follow-up tests suggested that the High-High

Table 9. Wolf and Smith motivation-contrasting group test performance descriptive statistics across test conditions

		Consequences		No Consequences	
Group	n	Mean	SD	Mean	SD
Low-Low	51	23.4	4.65	22.1	4.74
High-Low	28	27.1	6.21	24.6	4.51
Low-High	32	25.8	4.44	23.8	4.73
High-High	47	27.3	5.64	27.1	5.03

motivation group scored significantly higher than both the Low-Low and the Low-High contrasting-motivation groups. In the Wolf and Smith secondary analysis, it appears that individuals that are highly motivated in the consequential test conditions performed significantly better than individuals reporting low motivation in both test conditions. In the no-consequence test condition, the examinees reporting high motivation across test conditions significantly outperformed students in the Low-Low and the Low-High groups. These results again support the notion that motivation is related to higher performance.

Examinees with consistently high motivation do have higher test score averages than individuals with consistent self-reports of low motivation. In the consequence test condition, the pattern of scores was consistent with expectations that high consequence motivation would result in greater test performance. One additional analysis was conducted for both the current study data and the Wolf and Smith data to explore differential test performances across the experimental test conditions by motivation level. This analysis produced effect sizes comparing test performances across the two test conditions for each of the motivation contrasting-groups. Table 10 provides these effect sizes for both samples. While some of these effect sizes are quite large, their meaning is somewhat unclear. No apparent pattern emerged. It was expected that individuals reporting consistently high motivation would exhibit little change across test conditions. This expectation was borne out with the Wolf and Smith data but not with the current study. It was also expected that individuals with High consequence motivation and Low no-consequence motivation (High-Low) would produce the largest effect size across test

Table 10. Sundre and Wolf and Smith sample test performance effect sizes by motivation contrasting-groups.

Sample	Sundre	Wolf and Smith
Group	ES	ES
Low-Low	.98	.28
High-Low	.57	.47
Low-High	.00	.44
High-High	.44	.04

conditions. It may be that the method for creating contrasting groups was too arbitrary to provide meaningful comparisons. Perhaps the groups should have been formed with a criterion-referenced approach. Further research with additional samples is clearly necessary to elucidate the meaning of consequences and test performances for individuals with different motivation levels.

Results of hypothesis 5: Modality of assessment impacts motivation and performance. Effect sizes will be largest when comparing essay performance across consequence conditions.

This set of hypotheses introduced essay examinations to the design. The previous analyses provided support for the differential impact of consequences and motivation on multiple-choice test performances. These analyses replicate those conducted on the multiple-choice test with an essay format. Table 11 presents the descriptive statistics, dependent t-test result, and calculated effect size for motivation self-reports across the two essay consequential conditions. As with the selected-response examination, the

Table 11. Motivation descriptive statistics, t-test, and effect size comparing essay test conditions

Test condition	Mean	SD	t-test	df	p	ES
Essay Motivation						
Consequence	42.78	5.53	8.74	58	.000	1.59
Non-Consequence	30.93	9.58				

observed t-test result was highly significant. The effect size for this difference was also very large at 1.59. The effect size previously observed across consequential conditions for the multiple-choice motivation was .79. A parallel set of analyses was conducted for the motivation subtest scores Importance and Effort. These results are reported in Table 12. All dependent t-tests were significant, and the effect sizes were again frightfully large. It may be expected that ratings of importance for a test with no-consequence for poor performance would be significantly lower than the importance attributed to a test with consequence. However, if the Effort subscale is truly providing a measure of the level of effort examinees devote to assessment performance tasks in no-consequence

conditions, the ramifications for many testing programs are apparent and disquieting. Again, the linkage of motivation to performance must be made.

Table 12. Motivation subscore descriptive statistics, t-tests, and effect sizes comparing essay test conditions.

Test condition	Mean	SD	t-test	df	p	ES
Essay Motivation Importance						
Consequence	21.97	3.10	9.14	58	.000	1.57
Non-Consequence	15.57	5.07				
Effort						
Consequence	20.9	3.15	7.61	58	.000	1.32
Non-Consequence	15.4	5.14				

The real consideration for appropriate interpretation of test results must be related to the actual test performances. The next phase of analyses assessed essay performance differences across the two test conditions. Table 13 provides the results of these analyses. These results indicate statistically significant differences in essay performances across the test conditions. Higher performance was always observed in the Consequence condition. The effect size for this difference is 1.38, which is very high. These analyses

Table 13. Essay performance descriptive statistics, t-test, and effect size across test conditions.

Test condition	Mean	SD	t-test	df	p	ES
Essay						
Consequence	36.9	2.73	8.08	61	.000	1.38
Non-Consequence	28.6	9.37				

have shown consistently lower motivation and performances in the no-consequence test condition in every comparison to the consequence test condition. To provide a closer examination of the impact of test format modality on motivation, a set of analyses

comparing motivation self-reports within test conditions were performed. Table 14 presents the mean differences, dependent t-test results, and effect sizes comparing the multiple-choice with essay format motivation self-reports. In the consequential condition, all comparisons of motivation self-reports resulted in very small differences.

Table 14. Differences in test condition motivation self-reports by test modality.

Motivation Score	Multiple-Choice	Essay	t	p	ES
Test Condition: Consequence					
Total	43.0	42.8	.49	.63	.06
Importance	22.1	21.9	.56	.58	.04
Effort	20.9	20.8	.28	.78	.05
Test Condition: No-Consequence					
Total	37.0	30.9	6.68	.000	.71
Importance	18.3	15.6	6.10	.000	.60
Effort	18.6	15.3	6.07	.000	.71

No significant differences were observed, and all effect sizes were very close to zero. However, in the no-consequence experimental condition, the motivation self-reports resulted in significant differences. The motivation self-reports for the essay test modality are always lower, and in the no-consequences condition the effect sizes are all above .60. Thus, there is evidence that constructed-response examinations may result in lower motivation, effort, and performance in no-consequence testing conditions.

Summary

This study attempted to build upon a growing body of research exploring the impact of examinee motivation on test performances in conditions bearing no consequence for examinees. There have been a number of studies that have examined motivation self-reports and test performances within consequential and non-consequential conditions. The current study supported previous findings concerning increases in self-reports of motivation, and enhanced test performances in consequential testing conditions. The motivation subscores of Importance and Effort were found to be useful, and for some of the hypotheses, the two scores behaved differentially and in accordance

with expectations concerning the construct and the context. The Importance score consistently showed significant declines from consequential to non-consequential test conditions, as would be expected. The Effort scale did explain a bit more variance when correlated with test performance, and may prove to be an important tool for no-consequence test conditions.

The current study, replicating Wolf and Smith (1995) used an experimental no-consequence condition that could not simulate pervasive real-world testing conditions. It seems plausible that the effect sizes observed in this experimental condition with embedded classroom examinations assessing relevant, current class content may be a poor surrogate for the many no-consequence test conditions for which many examinees are 'volunteered.' We have not, as yet, captured or fully described the test score bias associated with those settings, though the current study and others like it have confirmed its presence and are attempting to estimate its magnitude.

Given the high reliability of the motivation scale and the two subscales, they may help to shed light on the true validity of many test performances gathered across international, national, district, school testing programs. Even in the experimental conditions used here, these scales helped to identify important examinee dispositions at the time of the testing. These scales helped to elucidate subtle but important differences. Each of the hypotheses tested resulted in findings supportive of the contention that consequences, examinee motivation, and task format count.

The effect sizes reported for the essay condition may be the most important results of the study. Previous research has suggested that examinees prefer the multiple-choice test format over essay. Some researchers have indicated that students enjoy alternative performance tasks; however, they did not report enjoying the challenge and demands associated with them. The novelty of these assessments will surely erode in time, and the research findings of this study and others indicate that the difficulty associated with these tasks will result in lower motivation and performances. The effect sizes observed for motivation and performance in the essay format compared to the multiple-choice were among the largest obtained in the study. Within the context of current testing practice, these study results may be of consequence.

References

Brown, S. M. and Walberg, H. J. (1993). Motivational effects on mathematics test scores of elementary-school students. *Journal of Educational Research, 86*(3), 133-136.

Burke, P. (1991). You can lead adolescents to a test but you can't make them try. (Contract No. OTA-H3-6110.0). Washington, D.C.: Office of Technology Assessment.

Ewell, P. T. (1991). To capture the ineffable: New Forms of assessment in higher education. *Review of Research in Education, 17*, 75-126.

Kim, J.-G., and McLean, J. E. (1995). The influence of examinee tesst-taking motivation in computerized adaptive testing. Paper presented at the annual meeting of the National Council on Measurement in Education. San Francisco, CA: April.

O'Neil, J. H. P., Sugrue, B., and Baker, E. L. (1995/1996). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. *Educational Assessment, 3*(2), 135-157.

Schmidt, M. J. & Ryan, A. M. (1992). Test-taking dispositions: A missing link? *Journal of Applied Psychology, 77*, (5), 629-637.

Sundre, D. L. (1997). Differential examinee motivation and validity: A dangerous combination. Paper presented at the annual meeting of the American Educational Research Association. Chicago, IL: April.

Wolf, L. F. and Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education,8* (3), 227-242.

Zeidner, M. (1993). Essay versus multiple-choice type classroom exams: The student's perspective. In B. Nevo & R. S. Jager (Eds.), Educational and psychological testing: The test taker's outlook (pp. 67-82). Toronto, Canada: Hogrefe & Huber.

Figure 1. Motivation Questionnaire Used by Wolf and Smith (1993).

Please circle one: This test counted. This test did not count.

Please think about the test that you just completed. Circle the number that best represents how you feel about each of the statements below.

1 = Strongly Disagree
2 = Disagree
3 = Neutral
4 = Agree
5 = Strongly Agree

1. Dong well on this test was important to me.
2. I am concerned about the score I receive on this test.
3. This was a very important test to me.
4. I gave my very best effort on this test.
5. I could have worked harder on this test. *
6. I did not give this test my full attention. *
7. I am eager to find out how well I did on this test.
8. I was highly motivated to do well on this test.

Figure 2. Motivation Questionnaire Used in this Study.

Student Opinion Survey

The results of this test

A = Counted for course credit
B = Did not count for course credit

Please think about the test that you just completed. Mark the answer that best represents how you feel about each of the statements below.

A = Strongly Disagree
B = Disagree
C = Neutral
D = Agree
E = Strongly Agree

1. Doing well on this test was important to me.
2. I engaged in good effort throughout this test.
3. I am not curious about how I did on this test relative to others. *
4. I am not concerned about the score I receive on this test. *
5. This was an important test to me.
6. I gave my best effort on this test.
7. While taking this test, I could have worked harder on it. *
8. I would like to know how well I did on this test.
9. I did not give this test my full attention while completing it. *
10. While taking this test, I was able to persist to completion of the task.

* Denotes items that are reversed prior to scoring.